

# Finite Countermodel Based Verification for Program Transformation

(A Case Study)

Alexei P. Lisitsa

Department of Computer Science,  
The University of Liverpool  
a.lisitsa@csc.liv.ac.uk

Andrei P. Nemytykh

Program Systems Institute,  
Russian Academy of Sciences\*  
nemytykh@math.botik.ru

Both automatic program verification and program transformation are based on program analysis. In the past decade a number of approaches using various automatic general-purpose program transformation techniques (partial deduction, specialization, supercompilation) for verification of unreachability properties of computing systems were introduced and demonstrated [10, 19, 30, 36]. On the other hand, the semantics based unfold-fold program transformation methods pose themselves diverse kinds of reachability tasks and try to solve them, aiming at improving the semantics tree of the program being transformed. That means some general-purpose verification methods may be used for strengthening program transformation techniques. This paper considers the question how finite countermodels for safety verification method [34] might be used in Turchin’s supercompilation method. We extract a number of supercompilation sub-algorithms trying to solve reachability problems and demonstrate use of an external model finder for solving some of the problems.

**Keywords:** program specialization, supercompilation, program analysis, program transformation, safety verification, finite countermodels

## 1 Introduction

A variety of semantic based program transformation techniques commonly called specialization aim at improving the properties of the programs based on available information on the context of use. The specialization techniques typically preserve the partial functions implemented by the programs, and given a cost model can be used for optimization. Specialization can also be used for verification of programs [10, 19, 30, 36] and for non-trivial transformations, such as compilation [9, 13, 51]. Essential for such applications is the interplay between semantic and syntactic levels. When the specialization is applied to a program  $P$  it may happen that semantic properties of  $P$  are transformed to simple *syntactic* properties of the residual program, which depending on the goals of transformation may be accounted in different ways.

It has been known for a while [10, 12, 29, 52, 53] that program transformation techniques, in particular program specialization, can be used to prove some properties of programs automatically. For example, if a program actually implements a constant function, sufficiently powerful and semantics preserving program transformation may reduce the program to a syntactically trivial “constant” program, pruning away unreachable branches and proving thereby the property. Viability of such an approach to verification has been demonstrated in previous work using supercompilation as a program transformation technique [35, 36] where it was applied to safety verification of program models of parameterized cache

---

\*The second author was supported by RFBR, research project No. 14-07-00133\_a, and Russian Academy of Sciences, research project No. 01201354590.

coherence protocols and Petri Nets models [26, 37]. Furthermore, the functional program modeling and supercompilation have been used to specify and verify cryptographic protocols, and in the case of insecure protocols a supercompiler was utilized in an interactive search for the attacks on the protocols [1, 2, 47].

Specialization can be used for analysis of other program properties, based on syntactic properties of the corresponding residual program. For example, any program can be seen as encoding a transition system of the parameterized program states and transitions and one may try to apply specialization to improve some properties of the transition graph (see [38]).

It is clear that any automated program analysis can be applied to program verification. On the other hand, many program analysis tasks, in particular those appearing in the context of program transformation techniques can be considered as verification problems.

In this paper we address the question of applications of a particular verification technique, *finite countermodel based verification (FCM)* [17, 33, 34, 49], within the context of *supercompilation*, a particular program transformation method. Supercompilation is a semantic based specialization method introduced by V. F. Turchin [53] which utilizes in particular unfold-fold based program transformation.

One of the main challenges in unfold-fold based program transformation techniques is to construct a *good* approximation of the *semantic* tree of the program  $P$  being transformed. We want to detect as many unreachable paths in the syntactic tree of  $P$  as possible and to prune away the dead paths. This syntactic tree is being stepwise unfolded from the program  $P$  definition. Thus, in essence, this challenge is a reachability problem. Supercompilation also poses itself other kind of reachability tasks and tries to solve them (see Section 5 for examples).

Finite countermodel method (FCM) for safety verification utilizes the principle of *reachability as derivability in the first-order logic* and reduces the task of safety verification, even for infinite state and parameterized systems, to the tasks of *disproving* first-order formulae, which are tackled then by available finite model finders. This principle, proposed initially for the verification of cryptographic protocols in [17, 49], has been later expanded to the larger classes of infinite state and parameterized verification problems, including safety for general term-rewriting systems [33]. The method has been shown to be *relatively complete* with respect to more widely known methods of *regular model checking* [34] and *regular tree model checking* [31] and generally it works when the safety can be demonstrated using *regular* invariants (see Section 2.1.1).

What makes the combination of supercompilation and the FCM method interesting and promising it is their somewhat complimentary features. On the one hand, the analysis mechanisms exploited within supercompilation do not cover *all* regular properties of parameterized program configurations (sets of states), so any help from the theoretically (relatively) complete FCM method could potentially be useful. On the other hand, supercompilation (as well as other specialization methods) sometimes is able to recognize and use non-regular properties of the program  $P$  (see Section 6 for examples) going beyond of what is possible by the FCM method.

So in short, the main idea of this paper is to explore the combination of FCM, theoretically complete for verification by regular invariants, with the mechanisms of supercompilation able to recognize and deal with non-regular properties.

This paper assumes that the reader has basic knowledge of concepts of functional programming, pattern matching, term rewriting systems, and program specialization. We also assume that the reader is familiar with the basics of the first-order logic.

## 2 Preliminaries

### 2.1 The Presentation Language

We present our program examples in a language  $\mathcal{L}$  which is a variant of a pseudocode for functional programs, while the supercompilation experiments with the programs were done in the strict functional programming language Refal [54].

The programs given below are written as *strict* (call by value) term rewriting systems based on pattern matching. The rewriting rules (also referred to as sentences) in the programs are ordered from top to bottom and they should be matched in this order.

The data set is a free monoid of concatenation (i.e., the concatenation is associative) with an additional unary constructor, which is denoted only with its parentheses (i.e., without a name). The colon sign denotes the concatenation. The constant  $\varepsilon$  is the unit of the concatenation and *may be omitted*, other constants  $c$  are characters. Let  $\mathcal{C}$  denote a set of the characters. The monoid  $\mathcal{D}$  of the data may be defined with the following grammar :

$$\mathcal{D} \ni d ::= \varepsilon \mid c \mid d_1 : d_2 \mid (d)$$

Thus a datum in  $\mathcal{D}$  is a finite sequence (including the empty sequence), which can be seen as a forest of *arbitrary* finite trees.

Let  $\mathcal{F} = \cup_i \mathcal{F}_i$  be a finite set of functional symbols, where  $\mathcal{F}_i$  is a set of functional symbols of arity  $i$ . Let  $v, f$  denote a variable and a function name, respectively. The monoid of the corresponding terms may be defined as follows:

$$t ::= \varepsilon \mid c \mid v \mid f(t_1, \dots, t_n) \mid t_1 : t_2 \mid (t), \text{ where } n \text{ is the arity of } f.$$

Let the denumerable variable set  $\mathcal{V}$  be disjointed in three sets  $\mathcal{V} = \mathcal{E} \cup \mathcal{S} \cup \mathcal{T}$ , where the names from  $\mathcal{E}$  are prefixed with 'e.', while the names from  $\mathcal{S}$  – with 's.' and the names from  $\mathcal{T}$  – with 't.'. s.variables range over characters, e.variables range over the whole data set  $\mathcal{D}$ , while a t.variable can take as value any character or any data surrounded by parentheses. For a term  $t$  we denote the set of all e-variables (t,s-variables) in  $t$  by  $\mathcal{E}(t)$  (respectively  $\mathcal{T}(t), \mathcal{S}(t)$ ).  $\mathcal{V}(t) = \mathcal{E}(t) \cup \mathcal{S}(t) \cup \mathcal{T}(t)$ .

Examples of the variables are s.r, t.F1, e.cls, e.memory. I.e., a variable name may be any identifier. We also use a syntactical sugar for representation of words (finite sequences of characters), so, for example, the list 'b':'a': $\varepsilon$  is shortened as 'ba' and 'aba':e.x denotes 'a':'b':'a':e.x.

We denote the monoid of terms by  $\mathcal{T}(\mathcal{C}, \mathcal{V}, \mathcal{F})$ . A term without function names is said to be *passive*. We denote the set of all passive terms by  $\mathcal{P}(\mathcal{C}, \mathcal{V})$ . Let  $\mathcal{G}(\mathcal{C}, \mathcal{F}) \subset \mathcal{T}(\mathcal{C}, \mathcal{V}, \mathcal{F})$  be the set of ground terms, i.e., terms without variables. Let  $\mathcal{O}(\mathcal{C}) \subset \mathcal{G}(\mathcal{C}, \mathcal{F})$  be the set of object terms, i.e., ground passive terms. Given a subset of the variables  $\mathcal{V}_1 = \mathcal{S}_1 \cup \mathcal{T}_1 \cup \mathcal{E}_1$  where  $\mathcal{S}_1 \subseteq \mathcal{S}$ ,  $\mathcal{T}_1 \subseteq \mathcal{T}$ ,  $\mathcal{E}_1 \subseteq \mathcal{E}$ , a substitution is a mapping  $\theta : \mathcal{V}_1 \rightarrow \mathcal{T}(\mathcal{C}, \mathcal{V}, \mathcal{F})$  such that  $\theta(\mathcal{S}_1) \subseteq \mathcal{C} \cup \mathcal{S}$  and  $\theta(\mathcal{T}_1) \subseteq \mathcal{C} \cup \mathcal{S} \cup \mathcal{T} \cup \{(t) \mid t \in \mathcal{T}(\mathcal{C}, \mathcal{V}, \mathcal{F})\}$ . A substitution can be extended to act on all terms homomorphically. A substitution is called *object* iff its range is a subset of  $\mathcal{O}(\mathcal{T})$ . We use notation  $s = t\theta$  for  $s = \theta(t)$ , call  $s$  an *instance* of  $t$  and denote this fact by  $s \ll t$ .

A program  $P$  is a pair  $\langle \tau, R \rangle$ , where  $\tau$  is a term called *initial*, and  $R$  is a finite list of rewriting rules of the form  $f(p_1, \dots, p_k) = r$ , where  $f \in \mathcal{F}_k$ , for each  $(1 \leq i \leq k)$ ,  $p_i$  is a passive term,  $r$  is a term containing the function names only from  $R$ ,  $\mathcal{V}(r) \subseteq \mathcal{V}(f(p_1, \dots, p_k))$ .

A program  $P = \langle \tau, R \rangle$  with  $R = \{l_i = r_i \mid i = 1 \dots n\}$  gives rise to a reachability relation  $\rightarrow_P$  on terms as follows. For  $t_1$  and  $t_2$  the term  $t_2$  is one-step  $P$ -reachable from  $t_1$  if and only if there exists a substitution  $\theta : \mathcal{V}(t_1) \rightarrow \mathcal{D}$  such that 1) there exists  $i : 1 \leq i \leq n$  such that for all  $j \in \mathbb{N}$ ,  $1 \leq j < i$ ,  $t_1 \theta$  does not match against  $l_j$  and it matches against  $l_i$ , and 2)  $t_2 = r_i \theta$ . In words,  $t_2$  is obtained from  $t_1$  by application of *the first, in the given order* applicable rewriting rule from  $R$ .

We denote by  $\Rightarrow_P$  a one-step reachability defined similarly to  $\rightarrow_P$  above, but omitting the clause “for all  $j \in \mathbb{N}$ ,  $1 \leq j < i$ ,  $t_1 \theta$  does not match against  $l_j$ ”. Thus  $t_1 \Rightarrow_P t_2$  iff  $t_2$  is obtained from  $t_1$  by application of any rule from  $R$ . It is obvious that  $\Rightarrow_P$  is an *overapproximation* of  $\rightarrow_P$ , that is  $\rightarrow_P \subseteq \Rightarrow_P$ . We denote the *reflexive, transitive closure* of  $\rightarrow_P$  and  $\Rightarrow_P$  by  $\rightarrow_P^*$  and  $\Rightarrow_P^*$ , respectively.

### 2.1.1 Term Automata, Regular Languages and Invariants

The following definition is an adaptation of the definition of *forest automata* from [4] to the specific kind of terms we introduced in Section 2.1.

**Definition 1** A term automaton over a finite set of characters  $\mathcal{C}$  and a finite set of functional symbols  $\mathcal{F} = \cup_i \mathcal{F}_i$  is a tuple  $\mathcal{A} = ((Q, e, *), \mathcal{C}, \mathcal{F}, \delta_{\mathcal{C}}, \Delta_{\mathcal{F}}, \delta_{()}, F \subseteq Q)$  where

- $Q$  is a finite set of states;
- $(Q, e, *)$  is a finite monoid;
- $\delta_{\mathcal{C}} : \mathcal{C} \rightarrow Q$ ;
- $\Delta_{\mathcal{F}} = \cup_i \{\delta_f : Q^i \rightarrow Q \mid f \in \mathcal{F}_i\}$  is a set of transition functions, one for each  $f \in \mathcal{F}$ ;
- $\delta_{()} : Q \rightarrow Q$  is a transition function for the unary constructor  $()$ ;
- $F$  is a set of accepting states;

For every ground term  $t \in \mathcal{T}(\mathcal{C}, \mathcal{F})$  the automaton assigns a value  $t^{\mathcal{A}} \in Q$  which is defined by induction:

- $\varepsilon^{\mathcal{A}} = e$ ;
- $c^{\mathcal{A}} = \delta_{\mathcal{C}}(c)$  for  $c \in \mathcal{C}$ ;
- $f(t_1, \dots, t_k)^{\mathcal{A}} = \delta_f(t_1^{\mathcal{A}}, \dots, t_k^{\mathcal{A}})$ ;
- $t_1 : t_2^{\mathcal{A}} = t_1^{\mathcal{A}} * t_2^{\mathcal{A}}$ ;
- $(t)^{\mathcal{A}} = \delta_{()}(t^{\mathcal{A}})$

A term  $t$  is accepted by the term automaton  $\mathcal{A}$  iff  $t^{\mathcal{A}} \in F$ . The term language  $L_{\mathcal{A}}$  of the term automaton  $\mathcal{A}$  is defined as  $L_{\mathcal{A}} = \{t \mid t^{\mathcal{A}} \in F\}$ . A term language  $L$  is called *regular* iff it is a term language  $L_{\mathcal{A}}$  of some term automaton  $\mathcal{A}$ .

A very general form of unreachability (safety) problem we consider in this paper can be specified as follows.

**Given:**  $R \subseteq \mathcal{T}(\mathcal{C}, \mathcal{F})$ , the set of *reachable* terms and  $T \subseteq \mathcal{T}(\mathcal{C}, \mathcal{F})$  the set of *target* terms;

**Question:** Is it true that  $R \cap T = \emptyset$ ?

We say that unreachability can be established by a *regular invariant* iff there is regular term language  $I$  such that  $R \subseteq I$  and  $I \cap T = \emptyset$ .

## 2.2 Examples

The infinite sequence **Fib** of Fibonacci words is defined recursively as

$$w_0 = b; w_1 = a; w_{i+2} = w_i w_{i+1};$$

and consists of the words:  $b, a, ba, aba, baaba, ababaaba, baabaababaaba, \dots$

**Example 1** The program  $\langle \tau, R \rangle$ , where  $\tau$  is  $\text{Fib}(\text{e.n})$  and  $R$  given below, computes the  $n$ -th pair of consecutive Fibonacci words, where  $n$  is given in the input argument (the  $\text{e.n}$  argument) in the unary notation. The result words are separated by the parenthesis constructors rather than the comma sign. For example,  $\text{Fib}(\text{'III'}) = (\text{'aba'}) : (\text{'baaba'})$ . Note that the right-hand side of the last rule uses associativity of the concatenation : the length value of  $\text{e.xs}$  is unknown and it may be greater than 1.

$\text{Fib}(\text{e.n}) = \text{F}(\text{e.n}, \text{'b'}, \text{'a'})$ ;

$\text{F}(\varepsilon, \text{e.xs}, \text{e.ys}) = (\text{e.xs}) : (\text{e.ys})$ ;

$\text{F}(\text{'I'} : \text{e.ns}, \text{e.xs}, \text{e.ys}) = \text{F}(\text{e.ns}, \text{e.ys}, \text{e.xs} : \text{e.ys})$ ;

The following example demonstrates the use of the associative constructor in the patterns (left-hand sides) of the first and second rules. The example program defines a predicate  $B$  testing: (1) whether or not the postfix of given Fibonacci word is  $\text{'bb'}$ ; (2) given two consecutive Fibonacci words, can the first of them end with  $\text{'b'}$ , while the second starts with  $\text{'b'}$ ? In the positive case the predicate value is  $\text{'F'}$ , otherwise it is  $\text{'T'}$ .

**Example 2**  $\tau$  is  $B(\text{Fib}(\text{e.n}))$  and  $R$  is from the previous example together with:

$B((\text{e.xs} : \text{'bb'}) : (\text{e.ys})) = \text{'F'}$ ;

$B((\text{e.xs} : \text{'b'}) : (\text{'b'} : \text{e.ys})) = \text{'F'}$ ;

$B((\text{e.xs}) : (\text{e.ys})) = \text{'T'}$ ;

### 2.2.1 Pattern Matching

Associativity of the concatenation creates an issue with the pattern matching, namely, given a term  $\tau$  and a rule  $(l = r) \in R$ , then there can be several substitutions matching  $\tau$  against  $l$ . An example is as follows:

**Example 3**  $\tau = f(\text{'abcabc'}, \text{'bc'})$  and  $l = f(\text{e.x} : \text{e.w} : \text{e.y}, \text{e.w})$ . There exist two substitutions matching these terms: the first one is  $\theta_1(\text{e.x}) = \text{'a'}$ ,  $\theta_1(\text{e.w}) = \text{'bc'}$ ,  $\theta_1(\text{e.y}) = \text{'abc'}$ , the second is  $\theta_2(\text{e.x}) = \text{'abca'}$ ,  $\theta_2(\text{e.w}) = \text{'bc'}$ ,  $\theta_2(\text{e.y}) = \varepsilon$ .

To make the pattern matching deterministic in the presentation language  $\mathcal{L}$ , we take the following decision arising from Markov's normal algorithms [40] and used in Refal [54]: (1) if there is more than one way of assigning values to the variables in the left-hand side of a rule in order to achieve matching, then we choose the one in which the leftmost  $\text{e}$ -variable takes the shortest value; (2) if such a choice still gives more than one substitution, then the chosen  $\text{e}$ -variable shortest value is fixed and the case (1) is applied to the leftmost  $\text{e}$ -variable from the  $\text{e}$ -variables excluding considered ones, and so on while the whole list of the  $\text{e}$ -variables in the left-hand side of the rule is not exhausted.

In the sequel we refer to this rule as Markov's rule and such a substitution as Markov's substitution on  $l$ , matching  $\tau$ .

**Example 4**  $\tau = f(\text{'abacad'})$  and  $l = f(\text{e.x} : \text{'a'} : \text{e.y} : \text{'a'} : \text{e.z})$ . There exist three substitutions matching the terms: the first is  $\theta_1(\text{e.x}) = \varepsilon$ ,  $\theta_1(\text{e.y}) = \text{'b'}$ ,  $\theta_1(\text{e.z}) = \text{'cad'}$ ; the second is  $\theta_2(\text{e.x}) = \varepsilon$ ,  $\theta_2(\text{e.y}) = \text{'bac'}$ ,  $\theta_2(\text{e.z}) = \text{'d'}$ ; the third is  $\theta_3(\text{e.x}) = \text{'ab'}$ ,  $\theta_3(\text{e.y}) = \text{'c'}$ ,  $\theta_3(\text{e.z}) = \text{'d'}$ .

The leftmost  $\text{e}$ -variable is  $\text{e.x}$ . Both in the first and the second substitutions the length of the  $\text{e.x}$ 's values is zero. The next leftmost  $\text{e}$ -variable is  $\text{e.y}$  and  $\text{length}(\text{'b'}) < \text{length}(\text{'bac'})$ . The first substitution meets Markov's rule.

Given a term of the form  $f(t_1, \dots, t_n)$  where for all  $(1 \leq i \leq n)$ ,  $t_i \in \mathcal{O}(\mathcal{C})$  and a term  $f(p_1, \dots, p_n)$  where all  $p_i$  are passive terms, the matching  $f(t_1, \dots, t_n)$  against  $f(p_1, \dots, p_n)$  can be viewed as solving the following system of equations in the free monoid of the object terms  $\mathcal{O}(\mathcal{C})$ .

$$\begin{cases} p_1 &= t_1 \\ &\dots \\ p_n &= t_n \end{cases}$$

We look for all values of the variables (i.e., substitutions  $\theta_i$ ) from  $\mathcal{V}(\mathfrak{f}(p_1, \dots, p_n))$  such that for each  $i$  and each  $(1 \leq j \leq n)$ ,  $\theta_i(p_j) = t_j$  and if the values' set is not empty we choose the only Markov's substitution, where Markov's rule acts on the pattern  $(p_1) \dots (p_n)$ . Note the patterns  $p_j$  may share variables and this equation system is equivalent to the following single equation  $(p_1) \dots (p_n) = (t_1) \dots (t_n)$ . This system has an important property: for all  $(1 \leq i \leq n)$   $t_i \in \mathcal{O}(\mathcal{C})$ . There is a simple algorithm solving the equation systems meeting this property.

### 2.3 On Supercompilation

In this paper we are interested in one particular approach in program transformation and specialization, known as supercompilation<sup>1</sup>. Supercompilation is a powerful semantic based program transformation technique [50, 53] having a long history well back to the 1960-70s, when it was proposed by V. Turchin. The main idea behind a supercompiler is to observe the behavior of a functional program  $p$  running on *partially* defined input with the aim to define a program, which would be equivalent to the original one (on the domain of the latter), but having improved properties. The supercompiler unfolds a potentially infinite tree of all possible computations of a parameterized program. In the process, it reduces the redundancy that could be present in the original program. It folds the tree into a finite graph of states and transitions between possible (parameterized) configurations of the computing system. And, finally, it analyses global properties of the graph and specializes this graph with respect to these properties (without additional unfolding steps). The resulting program definition is constructed solely based on the meta-interpretation of the source program rather than by a (step-by-step) transformation of the program.

The result of supercompilation may be a specialized version of the original program, taking into account the properties of partially known arguments, or just a re-formulated, and sometimes more efficient, equivalent program (on the domain of the original).

Turchin's ideas have been studied by a number of authors for a long time and have, to some extent, been brought to the algorithmic and implementation stage [46]. From the very beginning the development of supercompilation has been conducted mainly in the context of the programming language Refal [43, 44, 45, 54] based on syntax and semantics similar to that of our presentation language  $\mathcal{L}$ . A number of model supercompilers for subsets of functional languages based on Lisp data were implemented with the aim of formalizing some aspects of the supercompilation algorithms [24, 27, 42, 50]. The most advanced supercompiler for Refal is SCP4 [43, 44, 46].

## 3 Finite Countermodels and Program-State Reachability

Given a term  $t \in \mathcal{T}(\mathcal{C}, \mathcal{V}, \mathcal{F})$ , the set of instances of  $t\theta$  such that the substitution  $\theta$  is an object on  $\mathcal{V}(t)$ , ranging over  $\mathcal{O}(\mathcal{C})$ , is called *the state set* of  $t$ . Given a program  $P = \langle \tau, R \rangle$  in  $\mathcal{L}$  (see Section 2.1), the state set  $S_0$  of  $\tau$  is called *the initial state set* of  $P$ . The rewriting system  $R$  is able to evolve according to relation  $\rightarrow_P$ , starting from a fixed state  $s_0 \in S_0$  and producing some more states of  $P$ . Suppose that  $S_0$  is described by a predicate (characteristic function)  $\Sigma_0(\cdot)$  written in a logical language  $\mathcal{M}$ . Let  $\mathcal{B}$  be a formal theory defined in  $\mathcal{M}$  and  $\phi(\cdot)$  be a formula in  $\mathcal{M}$ , describing some state set of  $P$ . Assume that  $R$

---

<sup>1</sup>From *supervised compilation*.

satisfies the following: given two states  $s_0, s$  of  $P$ , if  $s$  is reachable via  $\rightarrow_P^*$  from  $s_0$  then  $\mathcal{B}, \phi(s_0) \vdash \phi(s)$ . Suppose that a formula  $\psi(\cdot)$  (hypothesis in  $\mathcal{B}$ ) specifies a set of bad states, whose reachability contradicts a safety property of the program  $P$ . Then refutation of  $\psi(s)$  (in the theory  $\mathcal{B} \wedge \phi(s_0)$ ) will mean the fact of safety of  $P$  – unreachability of the states satisfying the formula  $\psi(\cdot)$ . One may refute the hypothesis  $\psi(s)$  by producing a countermodel of the theory  $\mathcal{B} \wedge \phi(s) \rightarrow \psi(s)$ .

### 3.1 The Formal Theory of $\mathcal{D}$

Let us redefine the data monoid  $\mathcal{D}$  very slightly by providing an explicit name for the additional free unary operation. Henceforth, we assume that the characters set  $\mathcal{C}$  of the language  $\mathcal{L}$  is finite and  $\beta, \gamma \notin \mathcal{C}, : \notin \mathcal{C}$ . Let  $\beta$  stand for the unary operation. In this encoding the data set  $\mathcal{D}$  is redefined as follows:

$\mathcal{D} \ni d ::= \varepsilon \mid \gamma \mid d_1 : d_2 \mid \beta(d)$ , where  $\gamma$  ranges over  $\mathcal{C} \setminus \{\varepsilon\}$ .

Let  $\mathcal{C}$  be  $\{\varepsilon, 'a', 'b'\}$ , consider the following theory  $T_{\mathcal{D}}$  in the first-order predicate logic:

$$\begin{aligned} & \forall x, y, z. (x : y) : z = x : (y : z) \\ & \forall x. x : \varepsilon = x \\ & \forall x. \varepsilon : x = x \\ & (\neg(\varepsilon = 'a')) \wedge (\neg(\varepsilon = 'b')) \wedge (\neg('a' = 'b')) \\ & R(\varepsilon) \wedge R('a') \wedge R('b') \\ & \forall x. R(x) \rightarrow R(\beta(x)) \\ & \forall x, y. R(x) \wedge R(y) \rightarrow R(x : y) \end{aligned}$$

The first three axioms are the free monoid axioms: the first one expresses associativity of the concatenation, the second and third axioms say the constant  $\varepsilon$  is the identity element. The last three axioms axiomatize the unary predicate  $R(\cdot)$  reflecting the inductive definition of the data set.

The theory  $T_{\mathcal{D}}$  represents the data set  $\mathcal{D}$  as stated in the following proposition

**Proposition 1**  $d \in \mathcal{D} \Leftrightarrow T_{\mathcal{D}} \vdash R(d)$

## 4 Unfolding and $\mathcal{L}$ -Program-State Reachability

Let us briefly recall some basic concepts of program specialization which we need below. More details can be found in [42, 44, 50, 51, 53].

Given a function call  $\text{st}_0 = \mathbf{f}_k(d_0)$ , where  $\mathbf{f}_k \in \mathcal{F}_k$ ,  $d_0 \in \mathcal{D}^k$ , the abstract  $\mathcal{L}$ -machine  $\text{Int}(\mathbf{p}, \cdot)$ , starting from the state  $\mathbf{f}_k(d_0)$  by matching  $\mathbf{f}_k(d_0)$  against the left-hand sides  $l_i$  of the rules defining  $\mathbf{f}_k$ , chooses a particular rule  $\rho_{i_0}$  of  $\mathbf{f}_k$  and constructs a next state based on the right-hand side of  $\rho_{i_0}$ . This matching algorithm can be seen as an algorithm solving the following equations  $l_i = \mathbf{f}_k(d_0)$ . That is to say, the matching chooses the Markov's substitution  $\sigma(\cdot) : \mathcal{V}(l_i) \rightarrow \mathcal{D}$  such that  $\sigma(l_i) = \mathbf{f}_k(d_0)$ , if such a mapping exists. Let  $\text{Step}(\text{st}_0)$  denote the result of applying an algorithm including the successful pattern matching and the replacement of  $\text{st}_0$  with  $\sigma(r_i)$ .

In general,  $\text{Int}(\mathbf{p}, \cdot)$  iterates the execution of  $\text{Step}(\text{st})$  when the state  $\text{st}$  is a configuration from the function stack top and the input data of the state  $\text{st}$  are completely known.

The unfolding algorithm approximates the semantic tree of  $\mathbf{p}$  by means of iterating a meta-extension  $\text{MStep}(\cdot)$  of  $\text{Step}(\cdot)$  in the case when the state  $\text{st}$  is partially unknown. The execution of  $\text{MStep}(\text{st})$  results in a tree whose branches correspond to the subsets of the input parameters values. The branchings are produced by a meta-extension of the matching, i.e., by an algorithm solving the equations  $l_i = \text{st}$

of a general form in the term monoid  $\mathcal{T}(\mathcal{C}, \mathcal{V}, \mathcal{F})$ . In the sequel we refer to this meta-extension as the extended pattern matching. In particular, the algorithm has to solve word equations: in general, this task is nontrivial (see [21, 25, 39, 48]), although there exist several simple algorithms for solving such equations when they are of some restricted forms [7, 8, 22].

#### 4.1 One-Step Unreachability

Taking into account the hardness of Makanin's algorithm [39] solving word equations of general forms, one may approximate the semantic tree as follows. Let a program rule  $l = r$  and a parameterized state  $st$  be given. Before executing the algorithm  $F$  for solving the equation  $l = st$ ,  $MStep$  tries to prove that this equation has no solutions, using the algorithm  $NoSol$ , not necessarily complete for word equations.

If  $NoSol$  does not finish its work in the given time limit, we say that  $NoSol$  fails in proving the unsatisfiability. In the fail case we just call  $F$ . If  $NoSol$  succeeds, then the program rule  $l = r$  being considered is unreachable from the parameterized state  $st$  and the tree branch corresponding to this rule is pruned away. If  $F$  succeeds, then it, like a Prolog interpreter, may return a simple narrowing of the parameters of  $st$  and Markov's substitution depending on the narrowed parameters and satisfying the equation  $l = st$ . This substitution allows us to proceed with the unfolding. It may happen though that the narrowing of the parameters is not expressible in the language  $\mathcal{U}$  describing the parameterized configurations of the program being transformed. In particular, the set of solutions of such an equation may require a recursion for its definition, while the language  $\mathcal{U}$  may lack recursion and iteration constructions. We now exemplify the situation. The finite countemodelling finding is used as an  $NoSol$  procedure.

**Example 5** Consider the following program  $p = \langle \tau, R \rangle$ ,  $\tau = f('a' : e.q, e.q : 'a')$  and  $R$  is

$f(e.x, e.x) = 'T';$   
 $f(e.x, e.y) = 'F' : (e.x) : (e.y);$

The extended pattern matching has to solve the following system of the parameterized equations:

$$\begin{cases} e.x = 'a' : e.q \\ e.x = e.q : 'a' \end{cases} \quad (*)$$

It is equivalent to the single relation  $\Phi(e.q)$  (equation) –  $'a' : e.q = e.q : 'a'$  imposed on the parameter  $e.q$ .<sup>2</sup>

At the first glance,  $\Phi(e.q)$  must be the predicate labeling the first branch coming out of the semantics tree root, while the second branch must be labeled by its negation  $\neg\Phi(e.q)$ .  $\Phi(e.q)$  narrows the range of  $e.q$ . But the problem is that  $\Phi(e.q)$  cannot be represented in the pattern language, using at most finitely many patterns to define the program result of the unfolding. Recursion should be used to check whether or not a given input data belongs to the truth set of  $\Phi(e.q)$ .

The  $e.x$ -variable multiplicity  $\mu_{e.x}(f(e.x, e.x)) > 1$  causes this problem: the system  $(*)$  implies an equation, where the parameter  $e.q$  plays the role of a variable and both sides of the equation contain  $e.q$ .

Let us replace the initial parameterized configuration in Example 5:

$$\tau = f('a' : e.q : 'a' : e.q : 'b', e.q : 'a' : e.q : 'b' : e.q)$$

---

<sup>2</sup>It is easy to see that its solution set is  $\{\theta_i(e.q) = 'a'^i \mid i \in \mathbb{N}\}$ .



Now the extended pattern matching has to solve the following equation

$$'a' : e.q : 'a' : e.q : 'b' = e.q : 'a' : e.q : 'b' : e.q$$

It is easy to see that this equation having variable  $e.q$  both in the left and right-hand sides is inconsistent in  $\mathcal{D}$ . Mace4 automated finite model finder by W. McCune [41] quickly recognizes this fact in the context of the first-order theory  $T_{\mathcal{D}}$  (see Section 3.1). I.e., Mace4 finds a finite countermodel of the following formula  $\exists e.q. ('a' : e.q : 'a' : e.q : 'b' = e.q : 'a' : e.q : 'b' : e.q)$  in the theory of  $\mathcal{D}$ . That is, unfolding the program being considered in the given context can prune away the first rule of  $R$  and result in the following program (which can be viewed as a tree):

$p_1 = \langle \tau_1, R_1 \rangle$ ,  $\tau_1 = f_1('a' : e.q : 'a' : e.q : 'b', e.q : 'a' : e.q : 'b' : e.q)$  and  $R_1$  is the only rule:  $f_1(e.x, e.y) = 'F' : (e.x) : (e.y)$ ;

Note that we did not construct any narrowing of the parameter  $e.q$  and the information on the property of  $e.q$  (i.e., the equation above is inconsistent) is lost. The constructed tree is an approximation of the semantic tree of  $\langle \tau, R \rangle$ , rather than the exact semantics tree. If the right-hand side of the remaining rule includes a function call then, in general, the lost information might be useful for further specialization. For example, it might be  $'F' : g((e.x) : (e.y))$ . The configuration  $\tau_1$  might be encountered in an internal vertex of the unfolding tree.

The example above demonstrates a potential feature of using a finite countermodel finder for improving the approximation of the semantics tree generated by the unfolding. An interface linking the supercompiler SCP4 [43, 44, 46] with Mace4 has been implemented. It allows formulating in Mace4 such a kind of unreachability problem and returning to SCP4 the result produced by Mace4 during a time limit indicated by a user.

At first glance, the construction given in Example 5 can be generalized as follows. Let a program  $P = \langle t, R \rangle$ ,  $t = f(u_1, \dots, u_k)$  and  $R$  – a function  $f$  definition below, where  $k \in \mathbb{N}$ ,  $f \in \mathcal{F}_k$  and for all  $(1 \leq i \leq k)$   $u_i \in \mathcal{P}(\mathcal{C}, \mathcal{V})$ , be given. Let  $\# \mathcal{V}(t) = m \in \mathbb{N}$  and  $\# \mathcal{V}(l_i) = s_i \in \mathbb{N}$ . Let the sets  $\mathcal{V}(t), \mathcal{V}(l_1), \dots, \mathcal{V}(l_n)$  be ordered. Let  $q_j$  stand for the  $j$ -th element of  $\mathcal{V}(t)$ , while  $w_{ij}$  stand for the  $j$ -th element of  $\mathcal{V}(l_i)$ .

$$\begin{cases} f(p_{11}, \dots, p_{1k}) & = & r_1 \\ & \dots & \\ f(p_{n1}, \dots, p_{nk}) & = & r_n \end{cases}$$

**Definition 2** Given  $i \in \mathbb{N}$ ,  $1 \leq i \leq n$ , the rule  $l_i = r_i$  of the function  $f$  is said to be one-step reachable from the term  $t$  if there exists a substitution  $\theta : \mathcal{V}(t) \rightarrow \mathcal{D}$  such that for all  $j \in \mathbb{N}$ ,  $1 \leq j < i$ ,  $t\theta$  does not match against  $l_j$  and it matches against  $l_i$ . A rule is said to be one-step unreachable if it is not one-step reachable.

Let  $i \in \mathbb{N}$ ,  $1 \leq i \leq n$ , be given. One may suspect that refuting the following formula, expressing reachability of a rule  $l_i = r_i$ , leads to proving that the rule of the function  $f$  above is one-step unreachable from the term  $t$ .

$$\begin{aligned} \exists e.v \exists q_1, \dots, q_m. ((e.v = (u_1) \dots (u_k)) \wedge (\forall w_{11}, \dots, w_{1s_1} \neg (e.v = (p_{11}) \dots (p_{1k}))) \wedge \\ \dots \\ (\forall w_{(i-1)1}, \dots, w_{(i-1)s_{(i-1)}} \neg (e.v = (p_{(i-1)1}) \dots (p_{(i-1)k}))) \wedge \\ (\exists w_{i1}, \dots, w_{is_i} (e.v = (p_{i1}) \dots (p_{ik})))) \end{aligned}$$

But the variables here are assumed to range over the data set  $\mathcal{D}$  only and so the naïve application of a generic finite model finder may lead to vacuous countremodels, refuting the formula on a domain of

elements unrelated to  $\mathcal{D}$ . One may still try to analyse such conditions automatically, possibly using alternating applications of the first-order model finder and a theorem prover. We will address this issue elsewhere.

Nevertheless we can overapproximate the reachability condition. Namely, we do not consider any rule excluding the current rule being explored. That is to say, we approximate the  $\mathcal{L}$ -pattern matching with the pattern matching used in non-deterministic term rewriting. The corresponding formula to be refuted is as follows:  $\exists q_1, \dots, q_m. \exists w_{i1}, \dots, w_{is_i}. (u_1) \dots (u_k) = (p_{i1}) \dots (p_{ik})$ .

So the refutation of this formula proves unreachability by non-deterministic rewriting and therefore original unreachability.

Concluding this section we emphasize that the worst-case time complexity of the program resulting in Example 5 is  $\mathcal{O}(1)$ , while the worst-case time complexity of the original program is  $\mathcal{O}(n)$ , where  $n$  is the input data size. Improving this complexity was possible by the use of Mace4, which eliminated a rule of the original program with a hidden loop over the input data.

## 5 Global Unreachability

Unlike the most known specialization techniques supercompilation may extend the domain of the partial function defined by the program being transformed. That makes supercompilation more flexible as compared with those methods. For example, supercompilation is able to improve the worst-case time complexity of some input programs, while partial evaluation cannot [23]. Other transformation techniques such as distillation [18] can also improve the worst-case time complexity of some programs. In this section we consider one of the supercompilation tools for such a kind of transformations assisted by the finite countermodel method.

### 5.1 Online Generated Program Output Formats

Given a program  $P = \langle t, R \rangle$  and a substitution  $\theta : \mathcal{V}(t) \rightarrow \mathcal{D}$ , by  $\llbracket t\theta \rrbracket$  we denote the result of a (finite) computation of  $t\theta$  according to the program  $P$ .

**Definition 3** Let a program  $P = \langle t, R \rangle$  be given. A term  $u \in \mathcal{P}(\mathcal{C}, \mathcal{V})$  is said to be an output format of the program  $P$  if for any substitution  $\theta : \mathcal{V}(t) \rightarrow \mathcal{D}$  there exists a substitution  $\eta : \mathcal{V}(u) \rightarrow \mathcal{D}$  such that  $u\eta = \llbracket t\theta \rrbracket$ . Let  $u_1$  and  $u_2$  be two output formats of  $P$ . If  $u_1 \ll u_2$ , then we say  $u_1$  lesser than  $u_2$ . If  $u_1$  lesser than any other output format of  $P$ , then  $u_1$  is said to be a minimal output format of  $P$ .

The minimal output format of a given program is not unique. Examples of the output formats are:  $e.x$  is an output format of any program;  $s.y$  is the only (modulo variable renaming) minimal output format of the program defined in Example 2; both  $s.z : e.x$  and  $s.z : e.x : e.y$  are minimal output formats of the program given in Example 5.

The tree  $T$  being stepwise generated by unfolding is potentially infinite. As a consequence it is an object to be somehow folded back into a finite graph representing the residual program  $Q$ . The folding algorithm works stepwise online, i.e., given an intermediate state of  $T$  the algorithm tries to fold a potentially infinite path in this intermediate state into a loop, using generalization of parameterized configurations (states) of the original program  $P$ . We omit the details of the generalization algorithm. Edges folding such paths are called *references*. Thus the intermediate state of  $T$  actually is a graph  $G$  rather than a tree (see Fig. 1).

Given a vertex (a parameterized state of  $P$ )  $v$  of  $G$ , if all references from the vertices on the paths originating in  $v$  are incoming in the vertices from the same path set, then the corresponding part of  $G$

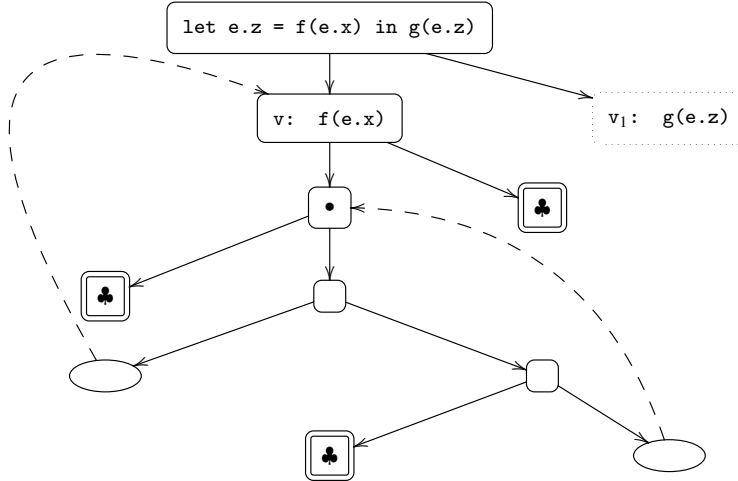


Figure 1: An intermediate state of an unfold-fold graph  $G$ : (1) the subgraph  $H$  rooted in  $v$  is self-sufficient, while (2) the subgraph rooted in  $\bullet$  is not. The configuration  $v_1$  still is not unfolded.

rooted in  $v$  is a self-sufficient (closed) subgraph. Such a vertex  $v$  is called the root of the subgraph. The root is a potential entry point into one of the residual functions (sub-programs), i.e., the root is an input format of a residual function  $H$ . Let such a root just be created by a step of the folding algorithm. The supercompiler SCP4 analyses the subgraph  $H$  and constructs its output format. The calls of the folding algorithm are stepwise interleaved with the calls of the unfolding algorithm. Hence  $G$  may include both some completely folded subgraphs and still non-unfolded parameterized configurations of  $P$ . Such configurations may include some calls of the residual function  $H$ . A non-trivial output format<sup>3</sup> of  $H$  restricts  $H$ 's image set. Therefore the information brought beyond  $H$ 's recursions (loops) might be used for specializing the function call (e.g.,  $g(e.z)$  in Fig. 1) using the call of  $H$  as an argument. That immediately implies the worst-case time complexity of the residual program  $Q$  may be reduced as compared with the worst-case time complexity of the original program  $P$ .

Note that the non-trivial output format may be a property of the original program  $P$  in the context of specialization (the initial configuration). In general, for example,  $P$  or its configuration being considered *per se* may have only the trivial output format. Thus the *online* generating of the output formats does matter.

The output format of the subgraph  $H$  is constructed by generalizing the formal exits from the recursions defining  $H$  (the double-boxing ♣ leaves in Fig. 1). Hence the lesser number of such exits the subgraph  $H$  includes the more specific output format of  $H$  may be constructed (see Definition 3 above). The recursion exits are presented as  $H$ 's edges belonging to the paths outgoing from the root  $v$ . Thus the following problem arises. Given a subgraph  $H$  representing a potential residual sub-program one has to prove unreachability of as many as possible syntactic recursion exits.

It is exactly the task we propose to delegate to a finite countermodel finder. Concrete executions of the finder may take too long times and even may not terminate. We suggest using the finder as explained in Section 4.1.

A formal theory for the image set (or its superset) of the partial function  $H$  needed to call the finder might be constructed by a compiler from the term rewriting language  $\mathcal{L}$  into the first-order logic lan-

<sup>3</sup>The trivial output formats are  $e.x$ ,  $e.x:e.y$ ,  $e.x:e.y:e.z$  and so on.

guage. The theory and a goal hypothesis to be refuted by the finder should somehow be based on the syntax of the subgraph corresponding to  $H$ .

For the reasons explained in Section 4.1 we can overapproximate the  $\mathcal{L}$ -unreachability with the non-deterministic term rewriting unreachability.

The simplest hypothesis is the assumption that several given syntactic exit-branches from the recursion defined by  $H$  are unreachable from the parameterized state in the root of the subgraph. Below we consider simple examples of such formal theories and goals.

## 5.2 Two Very Simple Output Formats

A very simple hypothesis is that the partial (sub)function being analyzed is *not* the empty partial function. I.e., at least one of its syntactic exits is reachable from the state in the root of the subgraph. Refuting this hypothesis means the subgraph root itself is unreachable from the root of the tree  $T$  and the branch leading in this subgraph root is dead. Any term in  $\mathcal{P}(\mathcal{C}, \mathcal{V})$  is an output format of the empty partial function.

If the analysis of the hypothesis above does not lead to the empty partial function, the following hypothesis could be made: the output format is a datum  $d \in \mathcal{D}$ . I.e., the (super)image set of a (sub)function being analyzed includes the single datum  $d$ . We might prove this fact if we are able to refute reachability of all syntactic exits excluding one, which returns the datum  $d$ .

Let us consider the program  $P$  given in Example 2. If the first two rules of the function  $B$  are unreachable from the initial configuration  $\tau = B(\text{Fib}(e.n))$ , then the minimal output format of  $P$  is 'T'. The supercompiler SCP4 is able to prove this fact by itself – without any call of a countermodel finder. This fact directly implies that none of the Fibonacci words contains 'bb' as a subword. Let us slightly change the predicate given in Example 2 as follows.

**Example 6** *Let us consider  $\tau$  to be  $A(\text{Fib}(e.n))$  and  $R$  to be from Example 1 together with:*

$$\begin{aligned} A( (e.xs : 'aaa') : (e.ys) ) &= 'F'; \\ A( (e.xs : 'aa') : ('a' : e.ys) ) &= 'F'; \\ A( (e.xs) : (e.ys) ) &= 'T'; \end{aligned}$$

SCP4 proves unreachability of the first two of the rules for  $A$  from the configuration  $A(\text{Fib}(e.n))$  and generates the minimal output format 'T'. That means none of the Fibonacci words contains 'aaa' as a subword. A more natural encoding of the same problem is presented in Example 7. For that encoding SCP4 fails to prove both properties of the Fibonacci words. It cannot recognize that the first rules of both  $A$  and  $B$  are unreachable from the corresponding initial configurations.

### Example 7

$$\begin{aligned} A( (e.xs) : (e.ys : 'aaa' : e.zs) ) &= 'F'; \\ A( (e.xs) : (e.ys) ) &= 'T'; \\ B( (e.xs) : (e.ys : 'bb' : e.zs) ) &= 'F'; \\ B( (e.xs) : (e.ys) ) &= 'T'; \end{aligned}$$

One may try to prove that those first two rules are unreachable from  $\tau$ , using the finite countermodel method (see Section 3). For the reasons explained in Section 4.1 we have to overapproximate the  $\mathcal{L}$ -reachability  $\rightarrow^*$  with the non-deterministic term rewriting reachability  $\Rightarrow^*$ . A first-order theory has to be generated, in which derivability is compatible with the overapproximated reachability condition in the program being considered.

Following [32] we here consider a simpler example of a first-order theory  $Fib_0$  demonstrating how to establish similar properties automatically using first-order theorem disproving by finite countermodels finding. The theory  $Fib_0$  is as follows:

$T_{\mathcal{D}}$   
 $K('b', 'a')$ .  
 $K(e.xs, e.ys) \rightarrow K(e.ys, e.xs : e.ys)$ .  
 $A(e.ys : 'aaa' : e.zs)$ .  
 $B(e.ys : 'bb' : e.zs)$ .

Here  $T_{\mathcal{D}}$  is the theory defined in Section 3.1, the meaning of the predicate  $K(e.xs, e.ys)$  is that  $e.xs$  and  $e.ys$  are two consecutive Fibonacci words. Negation of the last two axioms corresponds to the properties defined by the predicates  $A, B$  given in Example 7. Stepwise computation of a given Fibonacci word  $e.xs_0$  corresponds to stepwise derivability of  $\exists e.ys (K(e.xs_0, e.ys))$ . Mace4 is able to refute  $\exists e.xs \exists e.ys (K(e.xs, e.ys) \wedge B(e.xs))$  and  $\exists e.xs \exists e.ys (K(e.xs, e.ys) \wedge A(e.xs))$ , by finding countermodels  $M_1$  and  $M_2$  of sizes 5 and 11, respectively. Description of these models can be found in [32].

## 6 Regular Invariants and Beyond

The finite models produced above can be seen as compact representations of the *regular* invariants (see Section 2.1.1) and [33]) sufficient to prove safety, i.e., unreachability properties. The example above shows that enhancing of the supercompilation by the “regular verifying power” of FCM may be beneficial for producing non-trivial program transformations. What is interesting here is that the mechanisms for program analysis and transformation deployed within supercompilation do not cover all the regular power of FCM but may go beyond that.

Given a program rule  $l = r$ , obviously, using an  $e/t$ -variable  $v$  such that  $\mu_v(r) > 1$  may lead to one-step computing a non-regular formal language  $\mathcal{H} \subset \mathcal{D}$ . Such a language  $\mathcal{H}$  may also be generated by recursion. The following two examples deal with that case. The examples (being variations of the rules borrowed from [3]) define the empty partial function. The programs  $\langle \tau, R \rangle$  below never reach their exits from recursions. The exits are defined in the first two rules (in both programs). The first recursion given in the program  $F$  constructs two equal strings in the second and third arguments, using associativity of concatenation. Respectively the second recursion given in the program  $G$  constructs two equal binary trees, using the parenthesis constructor. The first arguments of the programs are the recursion depths. Evaluation of the programs generates respectively the following formal languages of terms:

$$H_f = f(K, \overbrace{'h' : 'A'}^n, \overbrace{'h' : 'A'}^n), H_g = g(K, \overbrace{(\overbrace{'h' : 'A'}^n)}^n, \overbrace{(\overbrace{'h' : 'A'}^n)}^n),$$

where  $K = ('b' \mid 'c')^m$  and  $m \in \mathbb{N}$ .

**Example 8** The program  $F$  is  $\langle \tau, R \rangle$ , where  $\tau$  is  $f(e.ps, 'A', 'A')$  and  $R$  is

$f(\varepsilon, 'h' : e.xs, 'A') = 'A';$   
 $f(\varepsilon, 'A', 'h' : e.ys) = 'A';$   
 $f('b' : e.ps, e.xs, e.ys) = f(e.ps, 'h' : e.xs, 'h' : e.ys);$   
 $f('c' : e.ps, 'h' : e.xs, 'h' : e.ys) = f(e.ps, e.xs, e.ys);$

**Example 9** The program  $G$  is  $\langle \tau, R \rangle$ , where  $\tau$  is  $g(e.ps, 'A', 'A')$  and  $R$  is

$g(\varepsilon, ('h' : e.xs), 'A') = 'A';$   
 $g(\varepsilon, 'A', ('h' : e.ys)) = 'A';$

```

g('b':e.ps, e.xs, e.ys) = g(e.ps, ('h':e.xs), ('h':e.ys));
g('c':e.ps, ('h':e.xs), ('h':e.ys)) = g(e.ps, e.xs, e.ys);

```

Proving the emptiness of the partial functions defined by  $F$  and  $G$  can be seen as safety verification, that is unreachability of the first two rules of  $F$  and  $G$  (i.e., the exits from the recursions). In [3] Y. Boichut and P.-C. Heam showed that this safety property cannot be proved by safety verification techniques using *regular invariants*. It means in particular that FCM won't help in proving that. On the other hand well-known specialization methods can prune away the recursion exits from program  $G$ . Indeed, the configuration sequence on the recursion path produced by the unfolding algorithm and outgoing from the initial configuration  $g(e.ps, 'A', 'A')$  is:  $g(e.ps, 'A', 'A')$ ,  $g(e.ps_1, ('h': 'A'), ('h': 'A'))$ ,  $g(e.ps_2, ('h': ('h': 'A')), ('h': ('h': 'A')))$ , ... Generalization algorithms based on the Higman-Kruskal relation [28] will construct one of the following configurations:  $g(e.ps, t.x, t.x)$ ,  $g(e.ps_1, ('h': t.x), ('h': t.x))$ ,  $g(e.ps_2, ('h': ('h': t.x)), ('h': ('h': t.x)))$ , ...

Note that  $t.x$  here denotes a standard variable like those occurring in Lisp-like languages to denote lists. Now obviously, the exit branches from the recursion will be pruned away from the unfolding tree.

The associative case used in  $F$  is more difficult. The supercompiler SCP4 recognizes the emptiness of both of the partial functions: supercompiling program  $G$  results in a message on the empty partial function, while supercompiling program  $F$  results in a program without *syntactic* exits from recursions.

Thus a finite countermodel finder and a specializer have incomparable power for verification and analysis and their joint use may be useful.

## 7 Conclusions and Future Work

External provers were used in automated program transformers including specializers for a long time. For instance, in 1988 Y. Futamura used such an external tool for proving some properties of parameterized configurations of programs being specialized [14, 15]. Moreover, given a program specializer written in a language  $\mathcal{U}$ ,  $\mathcal{U}$  may include non-trivial semantics mechanisms in itself. The mechanisms may allow us to implement non-trivial basic program analysis directly by means of the  $\mathcal{U}$ -semantics, i.e., through a *local* syntactical construction without intricate programming. Such mechanisms may be considered as external tools with respect to the specializer. Examples of such programming languages are Prolog and Refal [54]. For example, F. Fioravanti, A. Pettorossi and M. Proietti [10, 11, 12], as well as several other authors, develop an unfold-fold based transformation technique for constraint logic programs with negation, implementing their transformers using constraint logic programming.

The use of countermodels for the execution and analysis of logic programs has been considered in the paper [5]. It has been noticed that the failure of a query  $Q$  for a logic program  $P$  can be established by finding a countermodel for  $P \rightarrow Q$ . Furthermore a particular strategy using pre-interpretations (i.e., interpretations of the predicate symbols only, ignoring constructors and data) combined with the use of an abduction mechanism is proposed and compared with unrestricted search of countermodels. Such a technique can be adapted for term rewriting systems and functional languages.

The approach we presented in this paper is related also to the work on *abstract interpretations* [6, 30] and *regular types* [20]. The work [16] explicitly connects both areas and demonstrates the transformations of the set of regular type definitions corresponding to the finite tree automata, into a finite pre-interpretation for a logic program, which is then used for program analysis and verification. The core of the transformation is a determinization procedure for a non-deterministic tree automata. The difference with our approach, apart of obvious differences between logic and functional programming languages considered, is that [16] deals with specific approach for pre-interpretation building, while we abstract

away the details of the model building procedure, which is used as an oracle. Still after appropriate translations the approach of [16] can be used for the tasks considered in this paper and we plan to explore this issue in the future work.

In this paper we have shown that integrating a finite countermodel finder in a supercompiler may provide new features for non-trivial program transformations, which in turn may be used for non-trivial verification of safety properties of programs. In particular, global unreachability of some new kind of regular formal languages constructed over the system states sometimes may be recognized. Furthermore, Examples 8 and 9 demonstrate that Finite Countermodel Method (FCM) may be strengthened by supercompilation tools. As regards to this matter we would like to refer to an interesting example given by the researchers mentioned above, working in the context of Prolog [11]. They derive a one-counter machine from a constrained regular language specification. The corresponding residual program tests that a string of a given length  $n$  does not belong to the language  $\{ 'a^m' : 'b^n' \mid m = n \geq 0 \}$ .

Above we have described just the first steps and experiments in integrating FCM in a supercompiler. The examples given in Section 6 motivate future development of a compiler from a functional language (in our case, Refal) to a first-order logic language. The compiler should protect as many syntax properties of the program  $\langle \tau, R \rangle$  being compiled as possible. The *overapproximated reachability* of the  $\langle \tau, R \rangle$  states from  $\tau$  should correspond to *derivability* in the corresponding compiled program. We conclude with the following note: FCM may be used for recognizing unreachable intermediate subgraphs generated by supercompilation even if the subgraphs are not self-sufficient (see Section 5). In such cases we have to consider the external functions to be unknown.

## Acknowledgements

We are grateful to the reviewers of the paper for their generous and constructive comments, which allowed us to improve the presentation of this paper and gives us lines for future work.

## References

- [1] A. Ahmed, A. P. Lisitsa, and A. P. Nemytykh. Cryptographic protocol verification via supercompilation (A case study). In *VPT 2013*, volume 16 of *EPiC Series*, pages 16–29. EasyChair, 2013.
- [2] Abdulbasit M. Ahmed. Verification of cryptographic protocols via supercompilation. Master’s thesis, Department of Computer Science, University of Liverpool, 2008. 76pp, Available at URL <http://www.csc.liv.ac.uk/~alexai/A.Ahmed.dissertation.pdf>.
- [3] Y. Boichut and P.-C. Heam. A theoretical limit for safety verification techniques with regular fix-point computations. *Information Processing Letters*, 108(1):1–2, September 2008. doi:10.1016/j.ipl.2008.03.012.
- [4] M. Bojanczyk and I. Walukiewicz. Forest algebras. In J. Flum, E. Graedel, and T. Wilke, editors, *Logic and Automata: History and Perspectives*, Texts in Logic and Games, pages 107–132. Amsterdam University Press, October 2006.
- [5] M. Bruynooghe, H. Vandecasteele, Andre de Waal, and Marc Denecker. Detecting unsolvable queries for definite prolog programs. arXiv:cs/0003067, 2000.
- [6] P. Cousot and R. Cousot. Abstract interpretation: A unified lattice model for static analysis of programs by construction of approximation of fixpoints. In *Proc. of 4th ACM-SIGPLAN Symp. on Principles of Programming Languages (POPL’08)*, pages 281–292. ACM, 1977. doi:10.1145/512950.512973.
- [7] R. Dabrowski and W. Plandowski. Solving two-variable word equations. In *Automata, Languages and Programming*, volume 3142 of *LNCS*, pages 408–419. Springer Berlin Heidelberg, 2004. doi:10.1007/978-3-540-27836-8\_36.

- [8] V. Diekert. Makanin’s algorithm. In M. Lothaire, editor, *Algebraic Combinatorics on Words*, Chapter 12, pages 387–442. Cambridge University Press, 2002.
- [9] A. P. Ershov. On the partial computation principle. *Information Processing Letters*, 6(2):38–41, 1977.
- [10] F. Fioravanti, A. Pettorossi, and M. Proietti. Verifying ctl properties of infinite state systems by specializing constraint logic programs. In *the Proc. of VCL01*, volume DSSE-TR-2001-3 of *Tech. Rep.*, pages 85–96, UK, 2001. University of Southampton.
- [11] F. Fioravanti, A. Pettorossi, and M. Proietti. Transformation rules for locally stratified constraint logic programs. In K.-K. Lau and M. Bruynooghe, editors, *Program Development in Computational Logic*, volume 3049 of *LNCS*, pages 292–340. Springer, 2004.
- [12] F. Fioravanti, A. Pettorossi, M. Proietti, and V. Senni. Program specialization for verifying infinite state systems: An experimental evaluation. In M. Alpuente, editor, *LOPSTR 2010*, volume 6564 of *LNCS*, pages 164–183. Springer, 2011. doi:10.1007/978-3-642-20551-4\_11.
- [13] Y. Futamura. Partial evaluation of computing process an approach to a compiler-compiler. *Systems, Computers, Controls*, 2(5):45–50, 1971.
- [14] Y. Futamura, Z. Konishi, and R. Glück. Program transformation system based on generalized partial computation. *New Generation Computing*, 20(1):75–99, 2002. doi:10.1007/BF03037260.
- [15] Y. Futamura and K. Nogi. Generalized partial computation. In *the IFIP TC2 Workshop*, pages 133–151, Amsterdam, 1988. North-Holland Publishing Co.
- [16] John P. Gallagher and Kim S. Henriksen. Abstract domains based on regular types. In B. Demoen and V. Lifschitz, editors, *ICLP 2004*, volume 3132 of *LNCS*, pages 27–42. Springer, 2004. doi:10.1007/978-3-540-27775-0\_3.
- [17] J. Goubault-Larrecq. Finite models for formal security proofs. *Journal of Computer Security*, 6:1247–1299, 2010.
- [18] G. W. Hamilton. Extracting the essence of distillation. In *The Proc. of the 7-th International Andrei Ershov Memorial Conference: Perspectives of System Informatics*, volume 5947 of *LNCS*, pages 151–164. Springer Berlin Heidelberg, 2009. doi:10.1007/978-3-642-11486-1\_13.
- [19] G. W. Hamilton. Verifying temporal properties of reactive systems by transformation. *Electronic Proceedings in Theoretical Computer Science*, This Volume, 2015. The Proc. of the Third International Workshop on Verification and Program Transformation (VPT-2015).
- [20] E. K. Jackson, N. Bjørner, and W. Schulte. Canonical regular types. In *ICLP (Technical Communications)*, pages 73–83, 2011.
- [21] J. Jaffar. Minimal and complete word unification. *Journal of the ACM*, 37(1):47–85, Jan. 1990. doi:10.1145/78935.78938.
- [22] A. Jez. Recompression: a simple and powerful technique for word equations. In *(STACS 2013)*, volume 20 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 233–244. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2013.
- [23] N. D. Jones. *Computability and Complexity from a Programming Perspective*. The MIT Press, 2000.
- [24] P. A. Jonsson and J. Nordlander. Positive supercompilation for a higher order call-by-value language. *ACM SIGPLAN Notices*, 44(1):277–288, 2009. doi:10.1145/1480881.1480916.
- [25] Yu. I. Khmelevskii. Equations in free semigroups. (in Russian). In I. G. Petrovskii, editor, *Trudy Math. Inst. Steklov*, volume 107, 1971. English translation in: Proc. of Steklov Inst. Math., 107, Amer. Math. Soc., 1976.
- [26] A. V. Klimov. Solving coverability problem for monotonic counter systems by supercompilation. In *the Proc. of PSI’11*, volume 7162 of *LNCS*, pages 193–209, 2012. doi:10.1007/978-3-642-29709-0\_18.
- [27] I. Klyuchnikov. Supercompiler HOSC 1.5: homeomorphic embedding and generalization in a higher-order setting. Technical Report 62, Keldysh Institute of Applied Mathematics, Moscow, 2010.
- [28] J. B. Kruskal. Well-quasi-ordering, the tree theorem, and vazsonyi’s conjecture. *Trans. Amer. Math. Society*, 95:210–225, March 1960. doi:10.1090/S0002-9947-1960-0111704-1.



- [29] H. Lehmann and M. Leuschel. Inductive theorem proving by program specialisation: Generating proofs for Isabelle using Ecce. In *Proceedings of LOPSTR03*, volume 3018 of *LNCS*, pages 1–19, 2004. doi:10.1007/978-3-540-25938-1\_1.
- [30] M. Leuschel and T. Massart. Infinite state model checking by abstract interpretation and program specialisation. In *LOPSTR'99*, volume 1817 of *LNCS*, pages 63–82, 2000.
- [31] A. Lisitsa. Finite countermodels for safety verification of parameterized tree systems. CoRR, abs/1107.5142, 2011.
- [32] A. Lisitsa. Finite models for verification. a talk given in ENS Cachan, LSV, June 2012. <http://www.csc.liv.ac.uk/~alexei/countermodel/>.
- [33] A. Lisitsa. Finite models vs tree automata in safety verification. In *23rd International Conference on Rewriting Techniques and Applications RTA'2012*, pages 225–239, 2012.
- [34] A. Lisitsa. Finite reasons for safety. *Journal of Automated Reasoning*, 51(4):431–451, December 2013. doi:10.1007/s10817-013-9274-9.
- [35] A. P. Lisitsa and A. P. Nemytykh. Verification as parameterized testing (Experiments with the SCP4 supercompiler). *Programmirovaniye, (In Russian)*, 1:22–34, 2007. English translation in *J. Programming and Computer Software*, Vol. 33, No.1, pp: 14–23, 2007.
- [36] A. P. Lisitsa and A. P. Nemytykh. Reachability analysis in verification via supercompilation. *International Journal of Foundations of Computer Science*, 19(4):953–970, August 2008.
- [37] A. P. Lisitsa and A. P. Nemytykh. Solving coverability problems by supercompilation. Presentation on the Workshop on Reachability Problems - RP'08, 2008.
- [38] A. P. Lisitsa and A. P. Nemytykh. A note on program specialization. what can syntactical properties of residual programs reveal? In *VPT 2014*, volume 28 of *EPiC Series*, pages 52–65. EasyChair, 2014.
- [39] G. S. Makanin. The problem of solvability of equations in a free semigroup. (in Russian). *Matematicheskii Sbornik*, 103(2):147–236, 1977. English translation in: *Math. USSR-Sb.*, 32, pp: 129–198, 1977.
- [40] A. A. Markov. The theory of algorithms. *AMS Translations*, 2(15):1–14, 1960.
- [41] W. McCune. Prover9 and Mace4. [online]. <http://www.cs.unm.edu/~mccune/mace4/>.
- [42] N. Mitchell and C. Runciman. A supercompiler for core Haskell. In *Implementation and Application of Functional Languages*, volume 5083 of *LNCS*, pages 147–164. Springer-Verlag, 2008.
- [43] A. P. Nemytykh. The supercompiler Scp4: General structure. (extended abstract). In *the Proc. of PSI'03*, volume 2890 of *LNCS*, pages 162–170, 2003. doi:10.1007/978-3-540-39866-0\_18.
- [44] A. P. Nemytykh. *The Supercompiler SCP4: General Structure*. URSS, Moscow, 2007. (Book in Russian).
- [45] A. P. Nemytykh, V. A. Pinchuk, and V. F. Turchin. A self-applicable supercompiler. In *PEPM'96*, volume 1110 of *LNCS*, pages 322–337. Springer-Verlag, 1996.
- [46] A. P. Nemytykh and V. F. Turchin. The supercompiler Scp4: Sources, on-line demonstration. [online], 2000. <http://www.botik.ru/pub/local/scp/refal5/>.
- [47] Antonina Nepeivoda. Ping-pong protocols as prefix grammars and Turchin relation. In *VPT 2013*, volume 16 of *EPiC Series*, pages 74–87. EasyChair, 2013.
- [48] W. Plandowski. An efficient algorithm for solving word equations. In *Proceedings of the Thirty-eighth Annual ACM Symposium on Theory of Computing*, STOC '06, pages 467–476. ACM, 2006. doi:10.1145/1132516.1132584.
- [49] P. Selinger. Models for an adversary-centric protocol logic. *Electr. Notes Theor. Comput. Sci.*, 55(1), 2001.
- [50] M. H. Sørensen, R. Glück, and N. D. Jones. A positive supercompiler. *Journal of Functional Programming*, 6(6):811–838, 1996.
- [51] V. F. Turchin. The language Refal – the theory of compilation and metasystem analysis. Technical Report 20, Courant Institute of Mathematical Sciences, New York University, February 1980.

- [52] V. F. Turchin. The use of metasystem transition in theorem proving and program optimization. In *Proceedings of the 7th Colloquium on Automata, Languages and Programming*, volume 85 of *LNCS*, pages 645–657, 1980. doi:10.1007/3-540-10003-2\_105.
- [53] V. F. Turchin. The concept of a supercompiler. *ACM Transactions on Programming Languages and Systems*, 8(3):292–325, 1986. doi:10.1145/5956.5957.
- [54] V. F. Turchin. *Refal-5, Programming Guide and Reference Manual*. New England Publishing Co., Holyoke, Massachusetts, 1989. Electronic version: <http://www.botik.ru/pub/local/scp/refal5/>, 2000.